Chapter 5: The impact loss to authors and research

Michael Kurtz and Tim Brody

Introduction

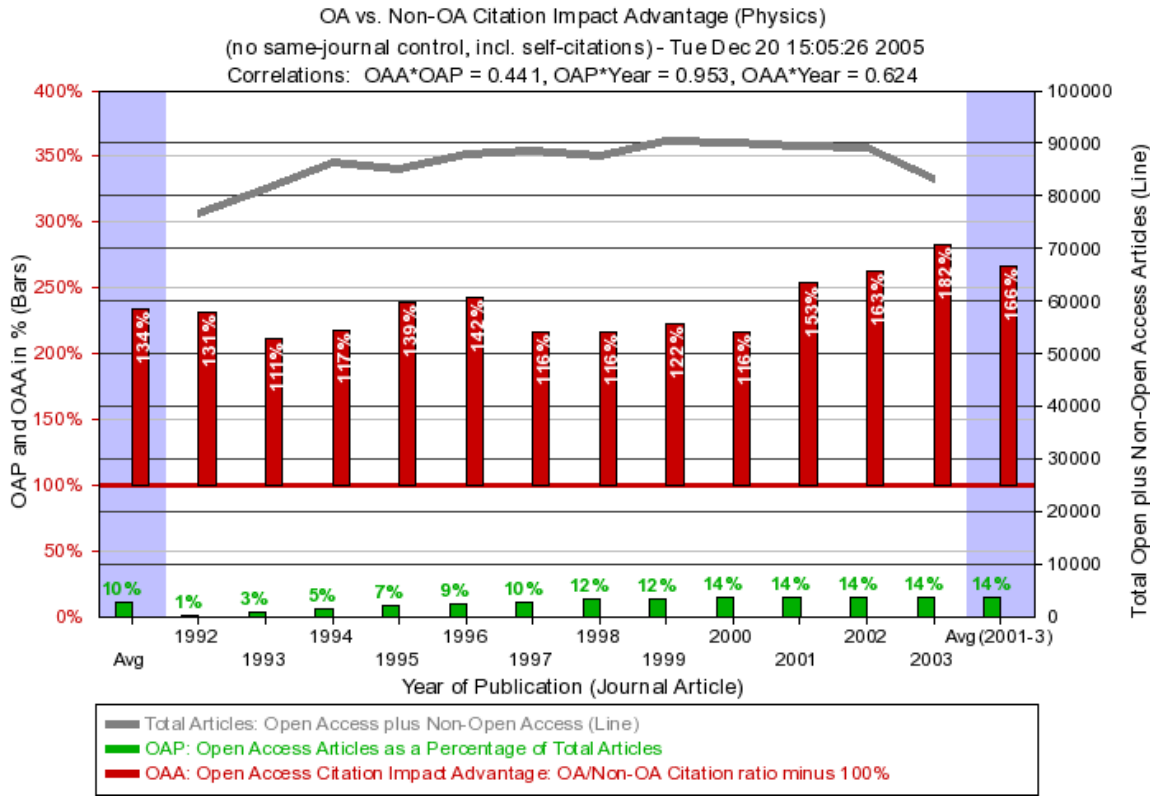
The history of scientific communication is one of increasing access. The Gutenberg press allowed rapid and relatively inexpensive reproduction of the printed word. The advent of postal services allowed for the distribution of papers across countries and around the world by airmail. Peer-reviewed journals created consistent collections of quality-controlled papers, distributed to a wider audience of subscribers. And, as the volume of journals increased, research libraries created collections of journals, catalogued, and made them accessible to patrons from the shelves. The web – and open access, OA – will allow anyone with an internet connection to access all the peer-reviewed literature anywhere, anytime. Increased accessibility of the peer-reviewed literature should allow that literature to have a greater impact on future research, which will improve the quality of that research. Those who invest in and benefit from primary research, including the general public, have an interest in improvements to the quality of that research. The authors of the peer-reviewed literature also have an interest in increasing its impact, since that impact, as traditionally measured using citation counts, is a major element in the way their work is evaluated.

Without debating the merits of evaluation by citation counting, this does provide a measurable (potential) benefit for authors that provide OA to their research papers. If OA increases citation impact – due to a greater number of scientists being able to access the paper – that presents a strong self-interest argument to encourage authors to go OA. It also hints at the extent to which restrictive access policies negatively affect research and its potential impact on future work.

Open Access articles have higher Citation Impact

Evidence for the greater citation impact of OA articles – compared to similar articles available only through subscription-based journals – has been shown by a number of studies. This greater citation impact of OA articles, as compared to similar subscription articles, is known as the citation advantage of OA. Lawrence (2001a) found a citation advantage for computer science articles freely available on the web, compared to articles available only through printed conference proceedings. An unpublished study (Brody, 2004a) performed jointly between the University of Southampton and the University de Quebec used the Thomson ISI *Science Citation Index* on CD-ROM to compare papers published in online peer-reviewed journals that were or were not available as e-prints in *arXiv*. Papers in arXiv are 'self-archived' by their authors – a version of the article, often a preprint, is deposited by the author. arXiv has become an indispensable tool for physics researchers, as it has grown to include almost all published literature in certain sub-fields, and provides early-day access to the published literature. The *Science Citation Index* contains references to papers published in some 5000 journals over 20 years, although this study is only relevant to an 11-year subset of that data (1992-2003 inclusive). Those papers found to be both in arXiv and the Science Citation Index received over double the number of citations compared to articles in the same ISI subject area, but not also available from arXiv. Figure 5.1 shows the proportion of papers in arXiv (OAP), the citation advantage (OAA), and the total papers broken down by the year an article was published. The

citation advantage is considerable, and increases noticeably in more recent years, probably because an increasing number of papers cite preprints that are only available via arXiv.



[Insert Figure 5.1]

Figure 5.1 Open Access Advantage for arXiv papers (based ISI citation data).

Hajjem 2005 has performed a similar study on other disciplines, using a web crawler to determine whether an article is 'OA'. Hajjem found an advantage between 25%-250%, depending on field and year.

The NASA Astrophysics Data Service (ADS) allows a comparison to be made between the number of citations to a journal, and the proportion of articles in that journal that have also been posted to arXiv. Querying for articles published in the Astrophysical Journal in 2003 finds 2592 articles cited 48388 times. Of those 2592, 1935 were found to have an equivalent in arXiv (75%), and accounted for 43411 of the 48388 total citations (90%). The 657 'non-arXived' articles received on average 7.58 citations each (4977/657). The other 1935 articles received on average 22.43 citations each (43411/1935). arXiv articles published in the Astrophysical Journal in 2003 received nearly three times the number of citations than non-arXiv articles published in the same journal (or a 196% OA citation advantage). Similar queries were performed for three other journals, and the results are given in Table 5.1.

	Nuclear Physics A	Physica	Astrophysical Journal	Physical Review D
Total Articles in 2003 (*cites*)	1134 *(2878)*	3920 *(3204)*	2592 *(48388)*	1990 *(24441)*
% in arXiv	32%	11%	75%	95%
citations	2590	1314	43411	23952
% OA Advantage	667%	548%	196%	181%

Table 5.1 OA advantage for four journals, based on data from the ADS (0% OAA = no advantage)

It is important to note that the OA advantage reported in Table 5.1 compares articles within a journal, rather than between journals. It is a competitive advantage for OA articles within a journal, compared to non-OA articles within the same journal. Comparing journals with each other, Pringle (2004) looked at the relative standing of OA journals in the total ISI journals index, and found little evidence for a citation advantage for them. Any difference that might be found would be more attributable to the editorial standards of quality, than the mode of access, and this underlines the limitations of comparing between journals, rather than between articles within a journal. However, that OA journals compare favourably against established, subscription-based journals is still quite surprising, given that the reputation – hence impact – of a journal takes time to build in a community

Kurtz (2005), studying astronomy, did not find any evidence that changing access from subscription to free increases the number of citations. Firstly, this confirms astronomy as a special case; it implies that:

> *"there is no significant population of astronomers who are both authors of major journal articles and who do not have 'sufficient' access to the core research literature. This also implies that increasing access above a 'sufficient' level has no influence on citation frequency."*

Secondly, it suggests that the wholesale shift to OA within the core literature of a discipline does not result in increased citation impact for that core. However, other advantages accrue to those disciplines that are wholly OA (see Odlyzko, this volume).

What causes the OA citation advantage?

Given that the OA advantage is at the article level, rather than at the journal level, what are its causes? It is clear that papers available through arXiv receive more citations on average than papers available only through subscription journals. Kurtz (2005), in a study based on seven leading astronomy journals, outlines three possible factors that could cause increased citation impact:

1. the advantage due to the article being openly (that is, freely) accessible (the 'open access advantage');
2. the advantage due to the article being accessible before its potential competitors, for example as a preprint on arXiv (the 'early access advantage');
3. possible bias in the OA sample of papers owing to authors tending to put on the web (for example, arXiv) more of what they consider to be their better papers ('self-selection bias').

A fourth possible factor is that inclusion in arXiv itself confers an advantage to a paper, because arXiv is indexed in various alerting and search services, and so raises the profile of the papers it contains (the 'arXiv advantage'). These four factors are discussed below.

The ‘open access advantage’ is the most difficult factor to test independently, since it demands comparisons between samples of articles that are similar in all other aspects – that is, articles that are provided by the same service(s) (to remove the ‘arXiv advantage’), on a comparable topic and at a comparable ‘quality’ level (to remove the ‘self-selection bias). One approach would be to conduct longitudinal analysis of citations to papers in a particular journal, as the proportion of OA papers increases.

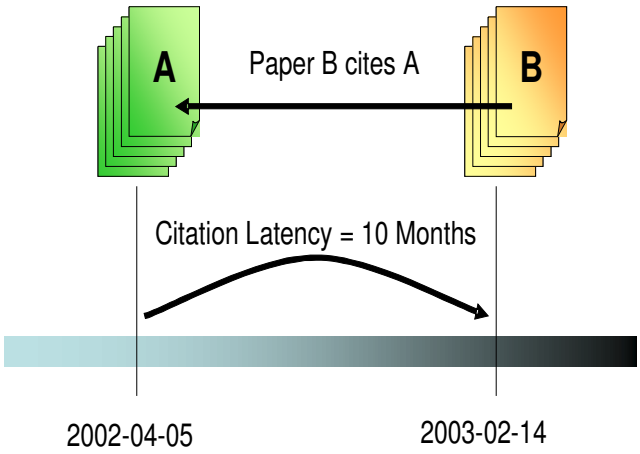
The ‘early access advantage’ occurs because the pre-print establishes priority and ‘presence’ in the literature, while the journal peer-review process provides the stamp of quality associated with a journal (and the associated impact the journal version affords). Henneken 2006 found that the ‘early access advantage’ is a permanent benefit – the higher the number of citations to an article, the more ‘pointers’ there are to that article, hence the more likely researchers will find, read, and cite it. Evidence showing that authors are increasingly citing articles in arXiv *before* those articles are published in a journal is discussed in the *citation latency* analysis that follows.

Kurtz (2005) provides strong evidence that author ‘self selection bias’ explains some of the strength of the citation advantage for articles in astronomy, where there is no OA effect (because the discipline operates as if it had OA, see above), and where the ‘early access advantage’ can be discounted. An author can’t know what the eventual citation impact of an article will be, but they may have a sense of its quality (hence likelihood to get cited), and opt not to put poorer work in arXiv. The obvious motivation would seem to be that authors are unwilling to place articles in the highly used arXiv service, but are still pressured to publish as much material in journals (as journal publication is recognised in formal research evaluation, and ‘grey’ literature – like arXiv – is not).

The ‘arXiv advantage’ occurs because arXiv is an invaluable resource for physicists: as well as providing full-texts, it provides alerting services that allow physics researchers to be emailed listings of new papers, simplifying discovery of new research. arXiv is also indexed by Google, as well as many research-specific search tools such as Elsevier’s *Scopus*, *OAIster*, and the new *ISI Web of Knowledge*. Another tool, *Citebase Search* (developed by the University of Southampton), provides citation navigation and search tools. The NASA *Astrophysics Data Service* (ADS) indexes arXiv as part of a wider collection and back-catalogue of Physics, Astrophysics and Astronomy papers. ADS provides email alerting services, similar to arXiv. All of these services increase the exposure – hence impact – of authors who place their papers in arXiv. This ‘arXiv advantage’, combined with ‘early access advantage’, provides a citation impact advantage to papers deposited in arXiv.

Citation latency

‘Citation latency’ is the time between a paper being published and it being cited. In the example in Figure 5.2, a paper ‘A’ is published on 5th April 2002. A subsequent paper ‘B’ is published on 14th February 2003 and cites paper ‘A’. The time difference between these two dates is 10 months, hence the citation latency for these pair of papers is 315 days. Because arXiv date stamps articles to the nearest day (and – being the web – articles are instantly accessible) it is possible to use an accuracy of days; something not possible in the on-paper era.



[Insert Figure 5.2]

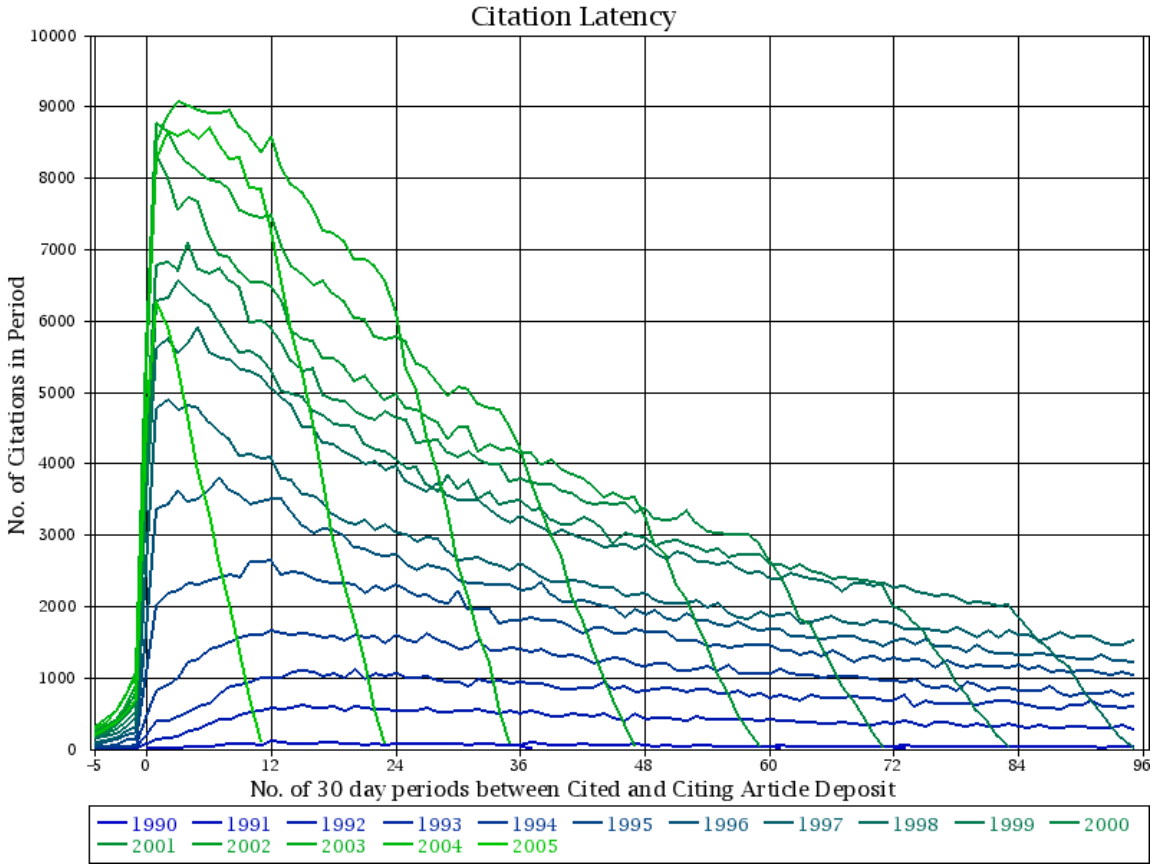
Figure 5.2 Definition of Citation Latency

Citation latency is measure of the efficiency of research communication. While research may be citable for a very long time - especially in natural fields, such as chemistry, where the properties and rules of the natural world don't change – most activity, and hence citations, tends towards more recent research. However, it has long been recognised that considerable differences exist between disciplines in this respect (Price, 1970).

As the volume and pace of research inexorably increases, so research publication must follow suit. It is no longer tenable to have research papers languishing in the publication cycle for years when scientific understanding can change dramatically in months. The web – and preprinting in particular – has a profound effect on the speed with which research results can be made public. The web allows the instantaneous transmission of information around the world, to any user with an internet connection, and yet few researchers are taking advantage of this to rapidly distribute (and establish priority on) the results of their research. Rapid pre-printing in arXiv demonstrates the effect that instant access to research results can have. Using arXiv as a case-study, it is possible to see this effect in action over the 15 years of arXiv's existence.

Open Access decreases citation latency

To analyse the effect that arXiv has had on physics communication the citation latency has been plotted for each year (Figure 5.3). All of the citations for each year are plotted according to the number of days between the citing and cited article being posted to the arXiv (Brody, Harnad and Carr, 2006). As the number of papers held in arXiv has been growing linearly since its inception, the number of citing papers (and hence citations) has grown linearly year on year. This is reflected in older years having a lower line. The oldest (hence lowest) years show a steady increase in the number of citations to a peak at around 12 months, then decreasing over time. This peak-point of citations has decreased each year until the most recent years where there is no apparent delay. This suggests that, as more physicists have deposited their papers in arXiv, so they have also increasingly cited arXiv papers and, with the near instant distribution nature of arXiv, so the peak-point of citation latency has reduced. As this data is based on *Citebase* (Brody, 2003) it only includes citations to arXiv articles, and biases those citations that include an arXiv identifier. That is, it doesn't necessarily follow that the cited 'half-life' has got shorter for articles in arXiv (the amount of time it takes for an article to receive 50% of the total citations it will ever receive).



[Insert Figure 5.3]

Figure 5.3 Citation latency measured across time.

Restrictive access policies restrict users

Kurtz (2004) compares the number of full-text accesses to a widely accessible journal in astronomy, the Astrophysical Journal, against publishers with fewer subscriptions and more restrictive access policies. Kurtz estimates the fraction of ADS users who do not have access to the Astrophysical Journal to be "a few percent or less." Hence it provides a 'virtual' OA baseline to test how many would-be users of other journals are denied due to toll-barriers. For the Astrophysical Journal, the fraction of total visits that included a request for the full-text was 63% (presumably the abstract was sufficient information for the 37% remainder). This was taken as a baseline, that is, it was assumed that 63% of visitors to other journals also wanted access to the full text. On this assumption, the Astrophysical Journal was compared to a number of other groups of journals (grouped by discipline/publisher) and, in the case of the group of journals with the most restrictive access policies, over half of all would-be users of the full text were denied access to it.

Correlation between use and citation impact

As journals have moved from print to the online medium, analysis of the *usage* of research articles has been made much easier. Web download analysis consists of counting the number of times users request the full-text of an article (sometimes augmented with an analysis of the

number of requests for pages about the article, such as an 'abstract' page). Similar to counting *citations*, the web download impact may be used as an indicator for the importance of that work. This argument is supported by the relationship between the number of times an article is cited (its citation impact) and the number of times it is downloaded (its download impact). Web downloads may even provide a better indication than citations for the usefulness of articles to a research community (Bollen, 2005).

A number of studies have calculated the correlation between the numbers of citations and downloads to individual articles. Perneger (2004) found a correlation (pearson's *r*) of 0.54 for the British Medical Journal between downloads from the journal site and citations from the *ISI Web of Science*. Moed (2005) found a correlation (spearman rank) of 0.35 for Tetrahedron Letters. Moed attempted to separate the mutual effects of citations on downloads and downloads on citations. Moed found a correlation of 0.11 between initial downloads and later citations, suggesting there is little predictive power in usage data. However, using *Citebase Search* (Brody, 2003) we have found the correlation between citations (from Citebase) and downloads (from the UK arXiv mirror) of r=0.44 (for the High Energy Physics sub-field, excluding first seven days of downloads). Restricting the period of downloads to three months (90 days) reduces the correlation to r=0.35, which appears to disagree with Moed's findings.

As has been shown by Kurtz (2004), restrictive access policies can result in a significant reduction in the number of accesses by users. Given the (at least partial) relationship between the number of accesses to full-texts and citation impact, reduced *download impact* may result in reduced *citation impact*.

Conclusions

Peer-reviewed journal articles also available as open access receive – on average – double the number of citations. This effect is easiest to measure early in the life of a journal article, as the preprint generates citations in addition to the journal article. The advantage is, however, sustained throughout the life of the article, as the more citations that point to an article, the higher the likelihood a researcher will navigate to it by following citation links (hence more reads, more impact, and more citations).

We have outlined a number of potential causes for the advantage conferred by open access: through rapid preprinting (which also leads to decreased *citation latency*), removing toll-barriers to authors (and researchers) without the journal subscription, depositing in high-profile open access services and a systematic effect of authors self-selecting higher quality material for author self-archiving.

The 'open access advantage' is a promising author-incentive to promote free access to the scholarly literature (with all the public benefit that comes with that), but establishing clear evidence that free access increases citation impact is beset with technical difficulties. So far, the evidence points towards greater access resulting in higher citation impact and, while some disciplines – like astronomy – are fortunate enough to already enjoy virtual 'open access' (by near-universal journal subscriptions), most disciplines and most countries will not be able to afford such widespread access. For authors in these fields, providing open access through open access journals and author self-archiving (while publishing through a high-quality journal), is the best way to maximise citation impact.